Anàlisi de dades mitjançant Business Intelligence: Un estudi de casos pràctics

Cristina Soler Bigorra

Resum—Aquest projecte explora l'aplicació de Business Intelligence (BI) i el seu potencial. Es divideix en dos parts: una primera on s'analitza un dataset amb un conjunt de dades reduïdes, i una segona on s'analitza un conjunt de dades més ampli. S'utilitzarà com a eina principal Power BI per realitzar visualitzacions de dades que permetin analitzar-les, i Python per l'aspecte més tècnic de programació. L'objectiu principal és extreure informació informació útil que pogués ajudar a prendre decisions informades. Durant el projecte, es crearan visualitzacions per identificar patrons de consum, predir tendències i optimitzar l'ús de les dades en diferents contextos. A través de l'anàlisi, es buscarà demostrar la rellevància del BI en la personalització i la millora de serveis.

Paraules clau—Business Intelligence, Power BI, Python, visualització de dades, anàlisi de consum, automatització, prediccions, datasets.

Abstract—This project explores the application of Business Intelligence (BI) and its potential. It is divided into two parts: the first part analyzes a dataset with a small set of data, and the second part analyzes a larger dataset. Power BI will be used as the main tool to create data visualizations that enable analysis, while Python will handle the more technical aspects of programming. The main objective is to extract useful information that could aid in making informed decisions. During the project, visualizations will be created to identify consumption patterns, predict trends, and optimize data usage in various contexts. Through the analysis, the goal is to demonstrate the relevance of BI in personalizing and improving services.

Index Terms—Business Intelligence, Power BI, Python, data visualization, consumption analysis, automation, predictions, datasets.

1 Introducció - Context del treball

El Business Intelligence (BI) consisteix en recopilar, analitzar i visualitzar dades per extreure'n conclusions i facilitar la presa de decisions. A través de diverses tecnologies, permet integrar i processar grans volums d'informació, identificar patrons, detectar anomalies i fins i tot predir tendències futures.

Un exemple clar d'ús del BI és el sector de l'entreteniment. Plataformes com Netflix, Amazon i Spotify l'utilitzen per analitzar les preferències dels usuaris i oferir recomanacions personalitzades. Per exemple, un usuari que compra a Amazon li poden recomenar productes simillars als que ha estat mirant anteriorment, així aumentant les possibilitats de que es realitzi una compra. A més, no només s'aplica en aquest àmbit, sinó que també en sectors com els supermercats o les botigues de roba, on ajuda a optimitzar estocs, personalitzar ofertes i millorar l'experiència del client.

Aquest Treball de Final de Grau (TFG) explorarà el concepte i les tecnologies del BI a través d'un cas pràctic. Per començar, s'analitzaran els tiquets de compra d'un únic client en un supermercat. A partir d'aquestes dades, s'utilitzarà Power BI, una eina de Microsoft per a la visualització de dades, amb l'objectiu d'identificar patrons de consum i extraure conclusions.

Un cop completada aquesta fase inicial, el treball

s'ampliarà amb un conjunt de dades més extens per aprofundir en les capacitats del BI. Això permetrà demostrar la importància del volum de dades en els processos analítics i confirmar com l'aplicació del BI pot ser un factor clau per a la personalització de serveis i la fidelització de clients en diferents sectors.

2 OBJECTIUS

L'objectiu d'aquest treball és principalment demostrar d'una forma pràctica la utilitat de la Business Intelligence. Bàsicament, veure de diferents maneres com aquesta pot ser útil tant com per un recull relativament petit de dades en ser un cas més casolà, com per obtenir decisions a partir d'una quantitat massiva de dades.

El treball tindrà com a objectiu primer analitzar un cas amb menys quantitat de dades, per així familiaritzar-se amb el Business Intelligence, i també tindrà una segona part on s'analitzarà un conjunt més gran de les dades, veient fins a quin nivell es pot analitzar amb aquesta eina. En aquest document m'he referit a aquestes parts com a primer cas i segon cas.

Es pot veure un exemple dels tiquets analitzats a l'apèndix A1. Exemple dades primer cas.

Uns objectius més concrets serien els següents:

- Identificar els patrons de les dades: buscar patrons dins de les dades, en el primer cas dels tiquets podrien ser els horaris de compra o els productes més buscats i tendències del client que s'està analitzant.

- Desenvolupar una visualització de dades: utilitzar Power BI per crear informes i dashboards que permetin visualitzar les dades de manera clara i eficaç.

- Extreure conclusions: a partir de les anàlisis elaborades, extreure conclusions com podrien ser, per exemple, recomenacions al client en el primer cas dels tiquets de compra.

- Realitzar prediccions: usar models d'Intel·ligència Artificial per poder fer prediccions partint de les dades.

2 METODOLOGIA

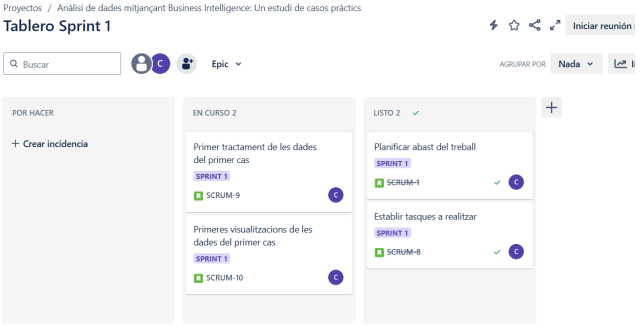
La metodologia que es farà servir estarà basada en Scrum. Tot i que aquesta està disenyada per a equips de treball, s'adaptarà i s'agafaran alguns dels seus conceptes per poder realitzar una versió individual. Així, es seguiran alguns dels seus principis per estructurar les tasques i posar fites.

El temps de treball es dividirà en Sprints, tenint en compte que el treball s'ha començat la setmana del 24 de febrer i acaba el 30 de juny. Per tant, consistirà de 18 setmanes de realització del TFG, y aquestes estaran dividides en Sprints de 3 setmanes. Per tant, el treball es farà en 6 Sprints.

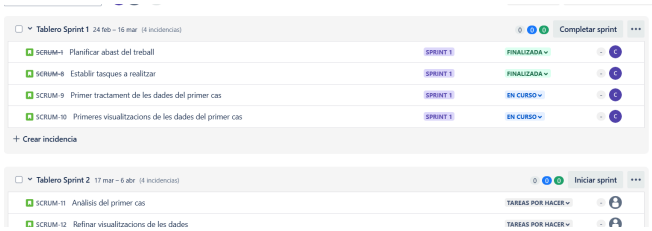
Primerament, es farà un recull de les tasques a realitzar, l'equivalent en Scrum del Product Backlog. Algunes d'aquestes seran la preparació de les dades o la creació d'un dashboard en Power BI. Aquestes tasques formaran la base de la planificació del meu treball. Un cop especificades aquestes tasques, es decidirà quines d'aquestes pertanyen al Sprint actual.

Al principi del Sprint, es revisaran les tasques que s'han de complir i quina prioritat tenen. Al ser un treball individual, no hi haurà reunions d'equip, sinó que cada setmana dedicaré mitja hora per mi mateixa a revisar el que he realitzat i el que falta encara per fer i quins problemes m'han sorgit. Aquesta serà una oportunitat més per ajustar la planificació al progrés de treball que es porti. Per acabar el Sprint, també es farà una reflexió similar, analitzant si he acabat el treball que m'havia proposat i si els resultats són adients o s'hauria de fer alguna cosa més per aconseguir-ho.

A més, cal especificar que es farà servir l'eina Jira per tenir un control del procés del treball. Allà s'organitzaran i s'actualitzara el progrés de les tasques, podent així monitorar l'estat del treball de final de recerca.



I en aquesta captura es pot veure com és el product backlog:



3 PLANIFICACIÓ

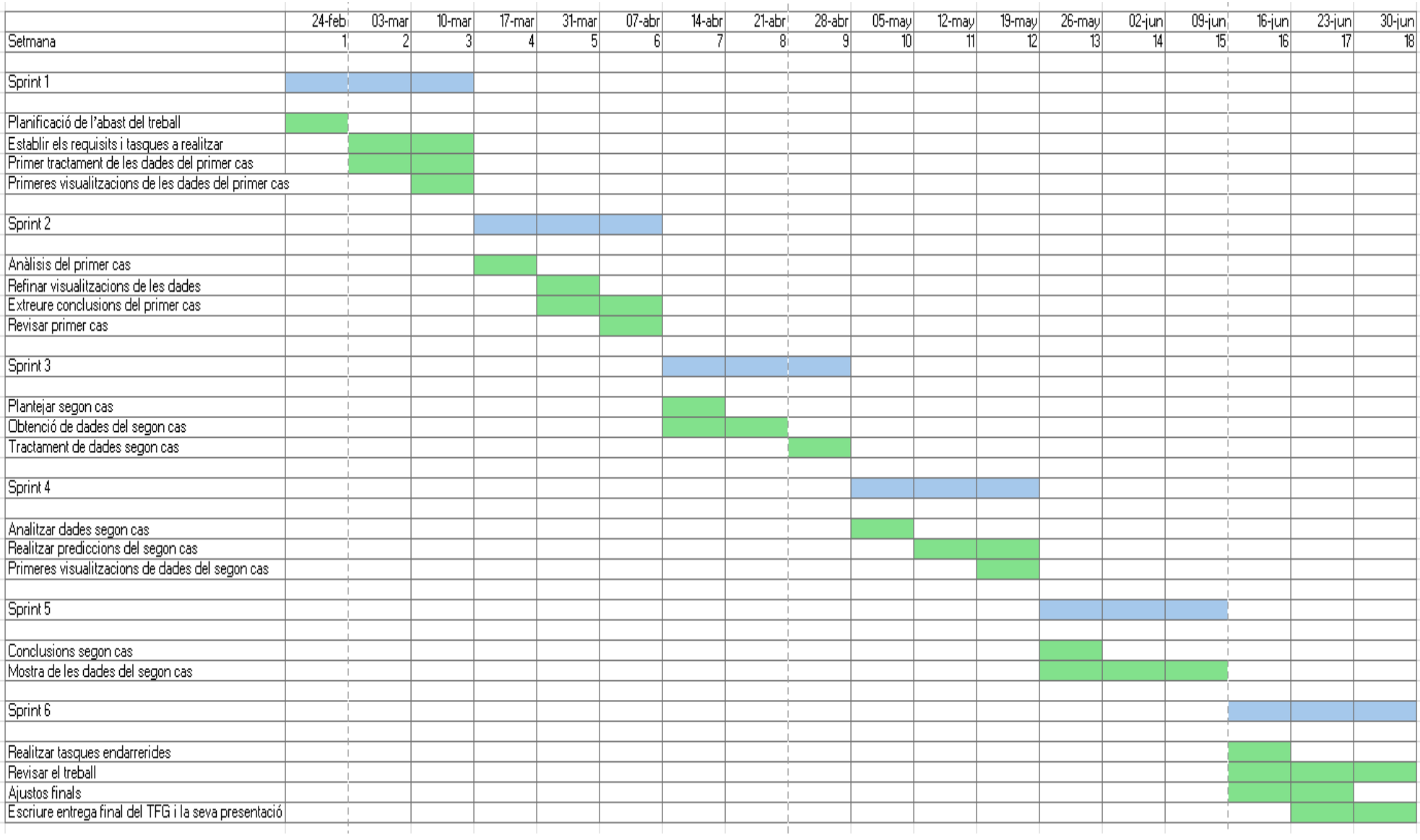
A continuació es mostra amb detall la planificació inicial, tot i que pot estar subjecte a canvis a mesura que progressa el treball, en cas de modificacions temporals depenent del seu progrés o si el treball acaba afegint noves funcionalitats i fent-se més extens.

La planificació estarà dividida en 6 Sprints, començant el 24 de febrer i acabant el 30 de juny. Els Sprints es poden veure a continuació a la taula 1.

Sprint	Durada	Objectius
Sprint 1	24 febrer - 16 març	Planificació de l'abast del treball Establir els requisits i tasques a realitzar Primer tractament de les dades del primer cas Primeres visualitzacions de les dades del primer cas
Sprint 2	17 març - 6 abril	Anàlisis del primer cas Refinar visualitzacions de les dades Extreure conclusions del primer cas Revisar el primer cas
Sprint 3	7 abril - 27 abril	Plantejar el segon cas Obtenció de dades del segon cas Tractament de dades segon cas
Sprint 4	28 abril - 18 maig	Analitzar dades segon cas Realitzar prediccions del segon cas Primeres visualitzacions de dades del segon cas
Sprint 5	18 maig - 8 juny	Conclusions segon cas Mostra de les dades del segon cas
Sprint 6	9 juny - 30 juny	Realitzar tasques endarrerides Revisar el treball Ajustos finals Escriure entrega final del TFG i la seva presentació

Taula 1: Planificació Sprints

A més, a la imatge 1 es pot veure el diagrama de Gantt dels Sprints.



Imatge 1: Diagrama de Gantt

4 DESENVOLUPAMENT

El desenvolupament del projecte s'ha centrat, en aquesta primera fase, en l'anàlisi de dades provinents dels tiquets de compra de Mercadona, corresponents a un únic client. Aquest conjunt de dades ha servit com a cas pràctic inicial per posar en marxa les tècniques i eines pròpies del Business Intelligence (BI), amb l'objectiu de convertir dades en brut en informació rellevant i visualment comprensible, obtenint gràfics que permetessin analitzar les dades.

4.1 Extracció i conversió de dades

Els tiquets es trobaven en format PDF, i per poder tractar les dades automàticament s'han passat a un format CSV. Per solucionar aquest repte, es va desenvolupar un script en Python que permet extreure la informació rellevant de cada tiquet i transformar-la en un fitxer estructurat per així utilitzar-lo com a dataset. Com es pot veure a la Imatge 2, el script fa ús de biblioteques com pdfplumber per extreure el text de cada pàgina del PDF, re per aplicar expressions regulars que localitzen patrons específics dins del text, i pandas per estructurar les dades en un DataFrame.

```python
import pandas as pd
import re
import os
import glob
import pdfplumber
from datetime import datetime

def extract_ticket_data(pdf_path):  1 usage
    data = []
    date = None
    total_import = None

    with pdfplumber.open(pdf_path) as pdf:
        for page in pdf.pages:
            text = page.extract_text()
            if text:
                lines = text.split('\n')
```

Imatge 2: Script extracció i conversió de dades

El procés comença recorrent tots els fitxers PDF d'una carpeta on estan guardats tots els tiquets, i per cada document:

- S'identifica la data de compra i es passa a un format de data.
- S'extreu informació dels productes, detectant dues estructures habituals:
  - Productes amb quantitat, preu unitari i total

1 FORMATGE TALLS LIGHT 2,32

  - Productes per pes, amb preu per quilo i total

1 BANANA
0,548 kg 1,45 €/kg 0,79

- S'extrau el total del tiquet
- Cada producte es guarda amb les columnes: data, nom del producte, quantitat/pes, preu unitari, preu total, i import total del tiquet

Finalment, les dades es guarden en un arxiu CSV anomenat ticket_dataset.csv, i es van afegint successivament les dades de cada PDF, amb el títol de les columnes només inserint-se la primera vegada. Això permet canviar el nombre de tiquets, per si per exemple s'havessin d'afegir més en un futur, que es puguin analitzar sense la necessitat de modificar el script.

Aquest script permet que les dades es guardin en un csv amb el format de la Imatge 3.

```
Fecha,Producto,Cantidad/Peso,Precio Unidad,Total,Importe Total Ticket
2024-01-27,BROU CASOLÀ DE PEIX,1,1.15,1.15,12.92
2024-01-27,BOSSA PLÀSTIC,1,0.15,0.15,12.92
2024-01-27,FLOR. BROCOLI,1,1.52,1.52,12.92
2024-01-27,CEBA TUB,1,1.85,1.85,12.92
2024-01-27,ALLIOLI TERRINA,2,1.35,2.7,12.92
2024-01-27,FLAN QUESO,1,2.0,2.0,12.92
2024-01-27,PROT. NORMAL,1,1.85,1.85,18.07
```

Imatge 3: CSV resultant de l'extracció de dades

4.2 Data curation

Un cop extretes les dades dels tiquets en format CSV, es va dur a terme un procés de neteja i estructuració per garantir la coherència i la qualitat de la informació abans de la seva anàlisi.

Aquest procés s'ha realitzat mitjançant un script en Python amb la llibreria pandas, com es pot veure a la Imatge 4.

```python
import pandas as pd

df = pd.read_csv("ticket_dataset.csv")

df['Producto'] = df['Producto'].str.strip()
df['Producto'] = df['Producto'].str.upper()
```

Imatge 4: Script data curation

S'han seguit els següents passos:

- Normalització dels noms de producte: es van eliminar espais innecessaris i es van convertir tots els noms a majúscules.
- Tractament de files incompletes: les files amb molts valors nuls s'han eliminat perquè no perjudiqui a l'anàlisi.

- Camps temporals: es crea un camp anomenat Mes_Compra, que indica en quin mes s'ha fet la compra, per així fer possible fer un seguiment temporal.

- Classificació de productes per categoria: amb un diccionari al script, es fa possible veure a quina categoria pertany cada producte, per així fer un anàlisi de quines categories compra més. Per exemple, FLAN QUESO pertany a LÁCTEOS Y HUEVOS.

El resultat final es va desar en un arxiu nou: tickets_final.csv, que inclou la data, el producte, la quantitat/pes, el preu, la categoria i el mes de compra.

El CSV resultant quedaria amb el format de la Imatge 5.

```
Fecha,Producto,Cantidad/Peso,Precio Unidad,Total,Mes_Compra,Categoria_Producto
2024-01-27,BROU CASOLÀ DE PEIX,1,1.15,1.15,1,CONSERVAS Y OTROS
2024-01-27,BOSSA PLÀSTIC,1,0.15,0.15,1,LIMPIEZA Y HOGAR
2024-01-27,FLOR. BROCOLI,1,1.52,1.52,1,FRUTA Y VERDURA
2024-01-27,CEBA TUB,1,1.85,1.85,1,FRUTA Y VERDURA
2024-01-27,ALLIOLI TERRINA,2,1.35,2.7,1,CHARCUTERÍA
2024-01-27,FLAN QUESO,1,2.0,2.0,1,LÁCTEOS Y HUEVOS
2024-01-27,PROT. NORMAL,1,1.85,1.85,1,CHARCUTERÍA
2024-01-27,SOPA POLLASTRE FIDEU,1,1.0,1.0,1,CONSERVAS Y OTROS
2024-01-27,PAPER HUMIT WC,3,1.55,4.65,1,LIMPIEZA Y HOGAR
2024-01-27,BLANQUEJADOR DENTIFR,1,3.95,3.95,1,LIMPIEZA Y HOGAR
```

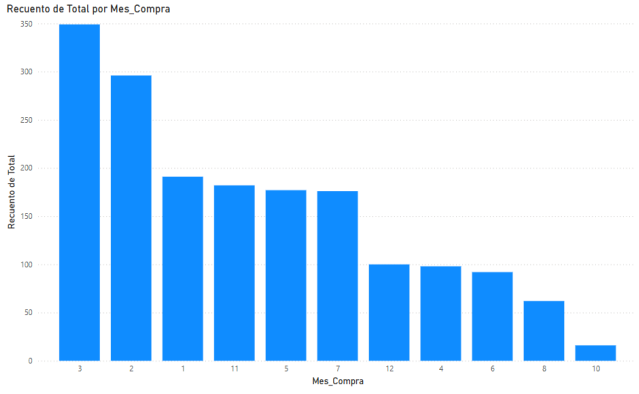
Imatge 5: CSV resultant de fer data curation

4.3 Visualització de dades amb Power BI

Amb les dades ja netes i estructurades, es va passar a la fase de crear visualitzacions utilitzant Power BI. Així es va poder analitzar el comportament de compra del client. Es van fer diversos gràfics, incluint boxplots i histogrames.

Algunes de les visualitzacions creades van ser:

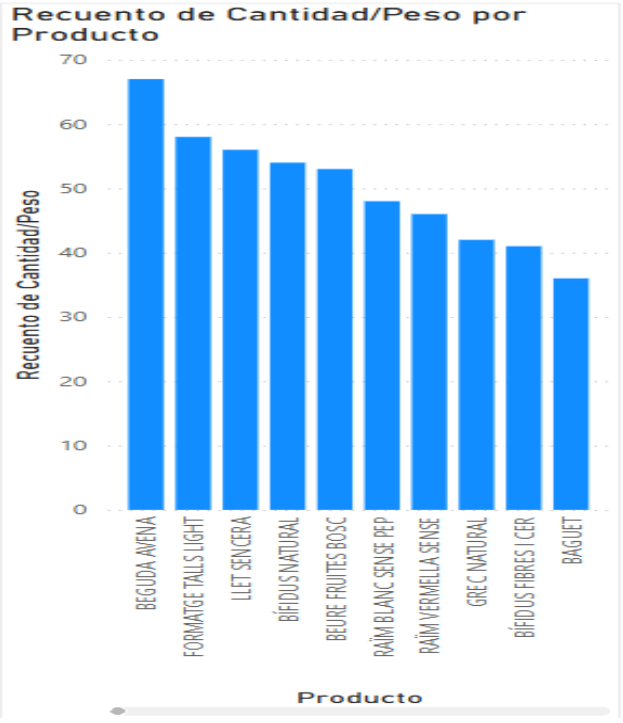
- Recompte de compres per cada mes (ordenat de més a menys).



Imatge 6: Gràfic de compres per mes

Es pot veure com a març, febrer hi va haver un número més elevat de compres. El tercer mes va ser gener amb no gaire diferència dels següents. El mes en el que menys es va comprar va ser octubre.

- Gràfics de barres amb els productes més comprats.



Imatge 7: Gràfic de compres per producte

També Power BI fa possible veure la informació com a una taula:

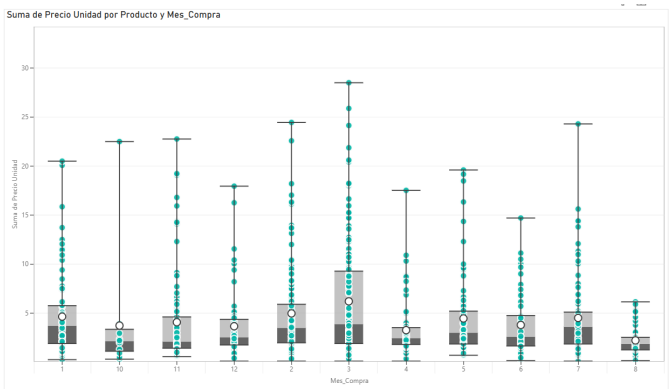
Producto	Recuento de Cantidad/Peso
BEGUDA AVENA	67
FORMATGE TALLS LIGHT	58
LLET SENCERA	56
BÍFIDUS NATURAL	54
BEURE FRUITES BOSC	53
RAÏM BLANC SENSE PEP	48
RAÏM VERMELLA SENSE	46
GREC NATURAL	42
BÍFIDUS FIBRES I CER	41
BAGUET	36
BOSSA PLÀSTIC	35
CERVESA VOLL-DAMM	33

Imatge 8: Taula de compres per producte

Tot i que el gràfic és més extens, aquesta part mostra els 10 productes més comprats. El primer seria la beguda avena amb 67 compres, després el formatge talls light amb 58, i després la llet sencera amb 56. En l'altre extrem també hi ha bastants productes que només s'han comprat una vegada, com podria ser wok pollastre.

- Boxplot de preus per observar la distribució i

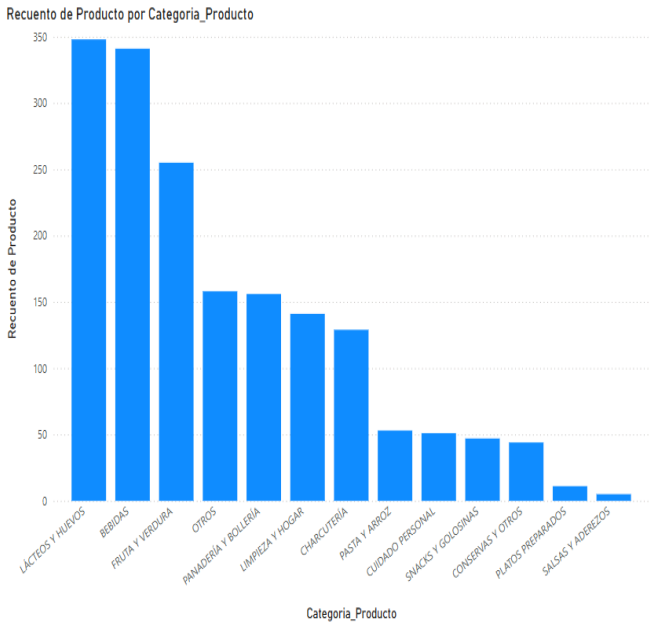
detectar possibles valors atípics, separant els productes per mes.



Imatge 9: Boxplot de preu del producte per mes

D'aquest boxplot es pot extreure diferent informació. Primer de tot es pot veure el producte més car que s'ha comprat cada mes. El més comú ha sigut l'oli verge, al ser el producte més car en quatre dels dotze mesos de l'any. A més es pot veure el que costa menys, per exemple la sal fina amb 0,4 euros a desembre. També es pot veure el canvi de preu dels productes segons el mes, per exemple l'oli verge al març costava 28,50 euros i al desembre 17,95 euros. Finalment, també es pot apreciar quina ha sigut la mitjana del preu dels productes comprats, per exemple a març era de 6,20 euros.

- Recompte de quins tipus de productes compra més el client:



Imatge 10: Gràfic de compra de tipus de productes

Es pot veure com els productes que compra més el client són làctics i ous, i seguidament begudes, mentre que els que compra menys són salses i plats preparats.

Aquestes visualitzacions ajuden a detectar fàcilment hàbits de consum com ara la compra recurrent de determinats productes, o en quins mesos es compra més.

4.4 Conclusions del primer cas

Aquest primer cas ha permès validar la metodologia aplicada i demostrar el potencial de la Business Intelligence fins i tot amb un volum de dades relativament reduït. A més el procés ha estat automatitzat, així que els scripts amb Python no haurien de ser modificats tot i que s'afegissin més dades.

A més, s'ha pogut comprovar com eines com Power BI, combinades amb scripts en Python per al tractament de dades, serveixen de forma molt afectiva per transformar dades en informació útil i visual. Aquest cas pràctic és un primer pas essencial cap a anàlisis més avançades, i servirà com a base pel segon cas d'aquest treball amb un conjunt de dades més extens.

Bibliografia

[1] "Introducción a Jira: guía completa para principiantes". Atlassian. Accedido el 3 de marzo de 2025. [En línea]. Disponible: https://www.atlassian.com/es/software/jira/guides/getting-started/introduction#what-is-jira-software

[2] "Introducción | Microsoft Power BI". Accedido el 3 de marzo de 2025. [En línea]. Disponible: https://www.microsoft.com/es-es/power-platform/products/power-bi/getting-started-with-power-bi

[3] R. Scrum, Agile Project Management: The Ultimate Step by Step Guide to Learn Agile Project Management to Complete Your Goals with Maximum of Results. Independently Publ., 2019.

[4] Kevin Stratvert. How to use Microsoft Power BI - Tutorial for Beginners. (4 de agosto de 2020). Accedido el 3 de marzo de 2025. [Video en línea]. Disponible: https://www.youtube.com/watch?v=TmhQCQr_DCA

[5] "Implementación de Soluciones de Business Intelligence (BI) con Python". 10Code Software-Entwicklung. Accedido el 3 de marzo de 2025. [En línea]. Disponible: https://10code.es/de/bi-python/

[6] "W3Schools.com". W3Schools Online Web Tutorials. Accedido el 3 de marzo de 2025. [En línea]. Disponible: https://www.w3schools.com/python/pandas/default.asp

[7] "W3Schools.com". W3Schools Online Web Tutorials. Accedido el 3 de marzo de 2025. [En línea]. Disponible: https://www.w3schools.com/python/default.asp

[8] "What are the phases/stages of a business intelligence project?" Le blog des meilleurs consultants indépendants I FocusTribes. Accedido el 3 de marzo de 2025. [En línea]. Disponible: https://blog.focustribes.com/en/business-intelligence-phases

[9] "A Gantt Chart Guide with Definitions & Examples". ProjectManager. Accedido el 3 de marzo de 2025. [En línea]. Disponible: https://www.projectmanager.com/guides/gantt-chart

[10] D. Lopez. "Cómo extraer datos de archivos PDF con Python". freeCodeCamp.org. Accedido el 3 de marzo de 2025. [En línea]. Disponible: https://www.freecodecamp.org/espanol/news/como-extraer-datos-de-archivos-en-pdf/

[11] freeCodeCamp.org. Machine Learning for Everybody – Full Course. (26 de septiembre de 2022). Accedido el 3 de marzo de 2025. [Video en línea]. Disponible: https://www.youtube.com/watch?v=i_LwzRVP7bg

---

· E-mail de contacte: cristinasolerb10@gmail.com

· Menció realitzada: Enginyeria del Software

· Treball tutoritzat per: Coen Antens

APÈNDIX

A1. EXEMPLE DADES PRIMER CAS

MERCADONA, S.A. A-46103834

C/ VILADOMAT 195
08205 SABADELL
TELÈFON: 937208041
11/02/2025 19:17 OP: 1656930
FACTURA SIMPLIFICADA: 3441-015-803829

	Descripció	P. Unit	Import
2	LLET SENCERA	0,94	1,88
1	BEGUDA AVENA		0,95
1	BOSSA PLÀSTIC		0,15
4	PA VIENA	0,40	1,60
1	RAÏM VERMELLA SESE		2,69
1	MAGDALENA SG 6UDS		2,55
1	ENCENEDOR CUINA WAVE		2,57
1	CUIDACOL NATURAL		2,46
1	BÍF FRUITES SIL		1,23
1	FORMATGE TALLS LIGHT		2,46
1	MANDARINA		
	1,404 kg	2,19 €/kg	3,07
1	PÀRQUING		0,00

ENTRADA 19:02 SORTIDA 19:17

TOTAL (€) 21,61
TARGETA BANCÀRIA 21,61

IVA	BASE IMPOSABLE (€)	QUOTA (€)
4%	13,26	0,53
10%	4,64	0,46
21%	2,25	0,47
TOTAL	**20,15**	**1,46**

MERCADONA 800 500 220

S'ADMETEN DEVOLUCIONS AMB TIQUET

DISPOSA DE 20 MINUTS
PER RETIRAR EL SEU VEHICLE